ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 2pSCb: Speech Intelligibility (Poster Session)

2pSCb3. The effects of temporal envelope confusion on listeners' phoneme and word recognition
Yingjiu Nie, Adam Svec, Peggy Nelson* and Benjamin Munson

*Corresponding author's address: Speech-Language-Hearing Sciences, University of Minnesota, 164 Pillsbury Dr SE, Minneapolis, MN 55455, peggynelson@umn.edu

Broadened auditory filters in listeners with hearing loss may result in listeners’ increased reliance on temporal envelope cues for understanding speech. Previous data have shown that background noise may affect hearing-impaired (HI) listeners by negatively affecting the temporal envelope cues in speech. The current study investigates additional HI listeners' understanding of vocoded spondees in the presence of fluctuating and stationary background noise. Stimuli were 8- and 32-channel noise vocoded double spondees, high-pass filtered at 1426 Hz. New data confirmed the previous finding that temporal envelope confusion in HI listeners resulted in speech understanding that is poorer in fluctuating noise (at a rate of 4Hz) than in stationary noise. Preliminary analysis suggests HI listeners experience significant envelope confusion for both 8- and 32-channel vocoded stimuli. Additional analysis of phoneme errors suggests that envelope confusion affects HI listeners’ perception of both consonants and vowels. Further analysis of j-factors will indicate the relationship of phoneme to whole word understanding for vocoded speech in noise. Results confirm the importance of temporal envelope cues for phoneme and syllable recognition for listeners with hearing loss. Work is supported by NIH DC008306 to PB Nelson

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Normal-hearing (NH) listeners show better speech understanding when speech is presented in fluctuating noise than in steady-state noise (Miller and Licklider, 1950), demonstrating the phenomenon of masking release. In contrast, reduced or absent release from masking has been demonstrated in hearing impaired (HI) listeners (Festen and Plomp, 1990; Bacon et al., 1998; Jin and Nelson, 2006; Lorenzi et al., 2006). A variety of explanations have been proposed for degraded masking release in listeners with HI. An audibility-based explanation suggests that elevated thresholds result in reduced audibility of speech in brief temporal windows when the noise is at lower levels (e.g., Takahashi and Bacon, 1992). However, investigators that have presented speech at suprathreshold levels have still observed reduced masking release in HI listeners (Takahashi and Bacon, 1992; Eisenberg et al., 1995; Bacon et al., 1998). A number of suprathreshold deficits in HI listeners have also been reported to account for the reduced/lack of masking release, including the shallower slope of recovery from forward masking (Rhebergen et al, 2006), degraded spectral resolution, presumably resulting from broadened peripheral auditory filters (Jin and Nelson 2006), and decreased ability to make use of temporal fine structure (TFS) (Buss et al., 2004; Lacher-Fougere and Demany, 2005; Lorenzi et al., 2006; Moore et al., 2006; Hopkins and Moore, 2007, 2010; Hopkins et al., 2008).

Kwon and Turner (2001) suggested that the amplitude modulation (AM) of the fluctuating noise interferes with the temporal envelope of speech, resulting in modulation interference and reducing masking release. They examined this hypothesized effect on consonant recognition in NH listeners using bandpass filtered vocoded speech samples /aCa/ in the presence of bandpass noise with variable spectral separations from the speech. The authors reported modulation interference when the noise was more spectrally separate from speech and masking release when the noise and speech fell in the same frequency region. However, it is not clear if and how this modulation interference affects masking release in HI listeners. A preliminary study (Nie and Nelson, 2012) using vocoded spondaic words suggested that the modulation interference may be more pronounced in mid- and high-frequency regions when the background noise is amplitude modulated at 4 Hz. This has been confirmed by recent findings of Jin, Nie and Nelson (2013).

Recently, Phatak and Grant (2012) suggested that AM rate of the background noise may affect vowel and consonant phoneme recognition to a different magnitude than observed for natural speech. They also suggested that modulation masking release observed in other studies (e.g., Jin and Nelson, 2006) likely reflected masking release for vowels as opposed to consonants. It is also hypothesized that individuals with a long history of hearing loss may have developed stronger reliance on contextual cues, which may facilitate their degraded speech understanding when the redundancy of speech cues is reduced. We explored this hypothesis by calculating the $j$ factor used by Boothroyd and Nittrouer (1988) as a means to evaluate the context effects at the word level.

The goal of the present project was to examine how modulation interference may affect masking release in HI when the TFS is removed and the spectral information is degraded. In addition, we investigated how varying spectral detail affects HI listeners’ masking release and modulation interference when the temporal envelopes are the major temporal cues. This investigation was motivated by the findings of Jin and Nelson (2006), whose results suggested that HI listeners may show improvements in masking release for natural speech if more detailed spectral information can be accessed. In addition, the present study looks into how the AM-rate of noise and varied amount of spectral details affect the vowel and consonant recognition in HI listeners, as well as use of context, when the temporal envelopes are the dominant temporal cues for real word recognition.

METHODS

Participants

NH listeners were recruited from undergraduate and graduate students ages 21 – 32 years. Their pure tone thresholds were 15 dB HL or less at audiometric frequencies of 250, 500, 1000, 2000, 4000, and 8000 Hz on the test ear. Two NH listeners have completed all the experimental sessions and three additional listeners are in progress.

HI listeners with pure tone thresholds 65 dB HL or less at the above audiometric frequencies were recruited. Table I shows the pure tone thresholds of the test ear of the HI listeners. The current analyses were based on the data collected from four HI listeners, including H06, H24, H28, and H31.
TABLE 1. Hearing impaired (HI) participants’ pure tone thresholds, ear tested, and age.

<table>
<thead>
<tr>
<th>Participants</th>
<th>Audiometric Frequency (Hz)</th>
<th>Ear tested</th>
<th>Age (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>250 500 1000 2000 3000 4000 6000 8000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H06</td>
<td>15 20 30 35 50 60 55 50</td>
<td>Left</td>
<td>58</td>
</tr>
<tr>
<td>H21</td>
<td>0 10 25 20 30 15 10</td>
<td>Right</td>
<td>26</td>
</tr>
<tr>
<td>H24</td>
<td>10 25 40 45 50 45 50 65</td>
<td>Left</td>
<td>50</td>
</tr>
<tr>
<td>H28</td>
<td>50 60 65 60 55 55 55 5</td>
<td>Right</td>
<td>24</td>
</tr>
<tr>
<td>H31</td>
<td>20 20 35 50 55 65 65 65</td>
<td>Left</td>
<td>72</td>
</tr>
<tr>
<td>H35</td>
<td>40 60 60 55 40 25 15 10</td>
<td>Right</td>
<td>25</td>
</tr>
</tbody>
</table>

Stimuli

218 spondee words spoken by a male talker were used in the experiments. The chance level for NH listeners to recognize the words was lower than 10% when these words were repeatedly presented (Davies-Venn et al, 2011). Two of the spondee words were randomly selected and spliced to form one presentation of four syllables (referred to as double-spondee). Background noise was generated with the same long-term average spectrum of the 218 spondee words of the male talker.

Vocoder Processing

The double-spondee words and white noise were filtered through either 8 or 32 filter bands whose cutoff frequencies were adopted from Fu and Nogaki (2004). Table 2 shows low and high cut-off frequencies of each of the band-pass filters at the two vocoder band resolutions (i.e., 8-band and 32-band).

TABLE 2. Low and high cutoff frequencies (in Hz) of the vocoder bands in the 8-band and 32-band conditions. The cutoff frequencies were adopted from Fu and Nogaki (2004).

<table>
<thead>
<tr>
<th>Band # (8-band)</th>
<th>1 2 3 4 5 6 7 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cutoff freq</td>
<td>200 359 591 931 1426 2149 3205 4748</td>
</tr>
<tr>
<td>High cutoff freq</td>
<td>359 591 931 1426 2149 3205 4748 7000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Band # (32-band)</th>
<th>1 2 3 4 5 6 7 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cutoff freq</td>
<td>200 233 272 312 359 408 464 524</td>
</tr>
<tr>
<td>High cutoff freq</td>
<td>233 272 312 359 408 464 524 591</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Band # (32-band)</th>
<th>9 10 11 12 13 14 15 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cutoff freq</td>
<td>591 664 745 833 931 1037 1155 1283</td>
</tr>
<tr>
<td>High cutoff freq</td>
<td>664 745 833 931 1037 1155 1283 1426</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Band # (32-band)</th>
<th>17 18 19 20 21 22 23 24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cutoff freq</td>
<td>1426 1581 1753 1941 2149 2376 2627 2902</td>
</tr>
<tr>
<td>High cutoff freq</td>
<td>1581 1753 1941 2149 2376 2627 2902 3205</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Band # (32-band)</th>
<th>25 26 27 28 29 30 31 32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low cutoff freq</td>
<td>3205 3537 3904 4305 4748 5233 5768 6354</td>
</tr>
<tr>
<td>High cutoff freq</td>
<td>3537 3904 4305 4748 5233 5768 6354 7000</td>
</tr>
</tbody>
</table>

The temporal envelopes of speech were extracted from individual vocoder bands with low cutoff frequencies at 1426 Hz or higher, by low-pass filtering the bands of speech through a 4th order Butterworth filter at a cutoff frequency of 50 Hz. The speech envelopes from these vocoder bands were half-wave rectified, and imposed on the white noise filtered through the same band. Each band carrying speech temporal envelopes was re-filtered through the same band-pass filter to eliminate the spectral spread resulting from imposing envelopes on noise. The individual bands of noise carrying speech envelope were added to formulate the high-pass vocoded-speech stimuli at a low cutoff frequency of 1426 Hz.

In noise conditions, the background noise was added to speech at the selected signal-to-noise (SNR) level prior to vocoding. The noise was either steady-state (SS) or sinusoidally amplitude modulated (SAM noise) at 4 Hz and 16 Hz at a 100% modulation depth. The mixture of speech and noise was modified through the vocoder processing.
Experimental Conditions

Three variables were examined in the present study, including: a) vocoder band resolution—8-band and 32-band; b) modulating conditions for the background noise—SS, SAM-4Hz, and SAM-16Hz; and c) SNR. Five SNR’s were investigated in both NH and HI listeners in the 8-band vocoder condition: -5, 0, 5, 10, and 13 dB. The same five SNR’s were examined for the 32-band vocoder condition for HI participants. Another five SNR’s were used for the NH participants in the 32-band vocoder condition: -5, -3, 0, +3, and 10 dB. Word recognition in quiet was also investigated for both 8-band and 32-band conditions. The total number of conditions was 32, including vocoder band resolution (2) X SNR (5) X modulation rate of noise (3) and two conditions for speech presented in the absence of noise (quiet) - one for each of the vocoder band conditions.

Procedure

The stimuli were processed by a customized Matlab code with a sampling rate of 44100 Hz and presented via a Lynx sound card controlled by a Pentium 4 computer. The speech stimuli were presented monaurally at 65 dB A in quiet for the high-pass stimuli. When background noise was present, the speech stimuli were presented at 65 dB A and the noise was presented at a level resulting in the nominal SNR. The SNR was determined by the difference in root-mean-square of the waveform in time domain between a double-spondee word and the background noise. For HI listeners, custom amplification following the half-gain rule was added to avoid the confounding factor of reduced audibility.

Participants completed 5-8-hours of practice prior to starting the experimental sessions. The practice was completed in two to four sessions (2-hours each) on different days. In the practice sessions, the double-spondee words were processed through vocoding with either a 16-band or 8-band vocoder band resolution. The vocoded speech was presented in full spectrum, low-pass, or high-pass filtered at 1426 Hz. The low cutoff frequency for the temporal envelope extraction was varied at 160, 50, 16, and 4 Hz. The speech material was presented in the absence of noise for all the practice sessions.

Listeners participated in four to eight experimental sessions (2-hours each) to complete the experimental conditions, each taking place on a different day. The 32 conditions were randomized for each listener, with one block of 20 double-spondee words presented for each condition. The 32 conditions were repeated in a different order, which, in addition to the first run, provided two scores per condition.

In both practice and experimental sessions, after the presentation of each double-spondee word, the listener was asked to respond by typing his/her answer into the experimental computer and then confirm the answer by clicking the mouse, following which the next word was presented. No feedback was provided. The listeners were allowed to take breaks as needed between the blocks of presentations. The participant’s responses were scored at either syllable or phoneme levels as described in the Results below.

RESULTS

Effect of Vocoder Band Resolution, AM rate, and SNR on Masking Release and Modulation Interference

Each double-spondee word was scored for the percent of the four single syllables correctly identified. Hence, two data points were obtained per test condition, each based on a list of 80 single words that formulated 20 double-spondee words. The percent correct was converted to rationalized arcsine transformed units (RAU). The following percentages are expressed in RAU, unless otherwise specified.

NH Listeners

Data collection is still in progress and further systematic analyses will be completed. Presently, two participants have completed the experimental sessions. The word recognition scores in quiet ranged between 52.3% and 74% for the 8-band condition and 92.3-120% (RAU) for the 32-band condition, suggesting a strong positive effect of increasing spectral detail in quiet.

Figure 1 shows NH listeners’ word recognition scores as a function of SNR for both 8-band (top panel) and 32-band (bottom panel) conditions in the presence of steady-state and SAM noise at 16 and 4 Hz. Linear least square
lines were fitted for different SAM rates of the noise. The data suggest better performance for the 32-band than the 8-band condition in noise, similar to the finding in quiet. The linear fits suggest that, in both 8-band and 32-band conditions, NH listeners scored higher when noise was amplitude modulated at 16 Hz than when the noise was either unmodulated or modulated at 4 Hz, especially at low SNRs.

**FIGURE 1.** Percent of syllables correctly identified as a function of SNR in NH listeners for the 8-band vocoder condition (top panel) and 32-band vocoder condition (bottom panel). Symbols depict individual listener's average scores (across 2 runs) for a given experimental condition. Linear least square lines are fitted for the data. SS: steady-state noise; SAM-16Hz: sinusoidally amplitude modulated noise at 16 Hz; and SAM-4Hz: sinusoidally amplitude modulated noise at 4 Hz.

**HI Listeners**

Word recognition scores in quiet ranged between 26% and 61.6% for the 8-band condition and 92.3-120% (RAU) for the 32-band condition. This suggests that HI listeners also showed an effect of improved performance with increasing spectral detail in quiet.

Figure 2 shows the word recognition score as a function of SNR for both 8-band (top panel) and 32-band (bottom panel) conditions in the presence of steady-state and SAM noise at 16 and 4 Hz. Three way repeated measures ANOVA (vocoder band resolution X SAM rate X SNR) performed with the data collected on four listeners showed no difference in performance between the two vocoder band conditions (i.e., 8-band and 32-band) ($F(1,3)=8.746$, $p>0.05$), suggesting no improvement with increasing spectral detail in noise. In addition, there was a significant
difference when noise was modulated at different rates \((F(2,6)=6.404, p<0.05)\), and a significant difference between SNR conditions \((F(2,6)=53.987, p<0.01)\).

Results suggest that the average score was highest when noise was modulated at 16 Hz, and lowest when the noise was modulated at 4 Hz. A significant interaction (SNR X SAM rate) was observed \((F(4, 12)=4.529, p<0.02)\). Preliminary investigation of this interaction revealed the highest average score for the SAM-16 Hz condition at the SNR of -5 dB, suggesting masking release at poor SNR’s. In contrast, at the more favorable SNR’s of 0 or 10 dB, the poorest performance was observed for the 4-Hz noise condition, suggesting modulation interference.

**FIGURE 2.** Percent of syllables correctly identified as a function of SNR in HI listeners for the 8-band vocoder condition (top panel) and 32-band vocoder condition (bottom panel). Symbols depict individual listener’s average scores (across 2 runs) for a given experimental condition. Linear least square lines are fitted for the data. SS: steady-state noise; SAM-16Hz: sinusoidally amplitude modulated noise at 16 Hz; and SAM-4Hz: sinusoidally amplitude modulated noise at 4 Hz.

**Effect of Modulation Rate on Consonant and Vowel Recognition**

The double-spondee words were scored for the percent correctly recognized at the phoneme level. Each spondee word consistently contained 2 vowels, while the number of consonants varied from 3 to 6. Preliminary phoneme scores were obtained for both NH and HI listeners in quiet and in the 8-band and 32-band conditions at selected SNR’s. The three SAM conditions were included for each of the band and SNR conditions. The SNR’s selected for analysis were determined by the criteria that no listener scored lower than 25% nor higher than 80% at a given SNR, to avoid ceiling and floor effects. In NH listeners, the 5 dB SNR condition was analyzed for the 8-band condition...
and SNR’s of -3 and 0 dB were included for the 32-band condition. In HI listeners, the 10 dB SNR was included in the analysis for the 8-band condition and the 5 dB SNR for the 32-band condition. The percent correct was converted to the RAU unless otherwise specified.

Presently, preliminary data analyses were performed on 2 NH listeners and 3 HI listeners (H06, H28, and H31). Paired t-tests were performed between the vowel and consonant recognition scores in NH listeners. The various background conditions were pooled. Overall, no difference has been observed in the recognition score between consonants and vowels (p=0.1249). Paired t-tests were performed for the HI listeners in the same way, showing that HI listeners’ consonant recognition was more negatively affected than vowel recognition (P<0.01). In addition, for both NH and HI listeners, both consonant and vowel recognition were equally affected by the modulation of the noise, regardless of whether the background noise was steady-state or amplitude modulated at 4 or 16 Hz. In other words, the consonant and vowel scores were equivalent in the NH group regardless of whether the noise was steady or amplitude modulated; whereas, consonant recognition remains poorer across all three noise conditions in HI listeners.

**j Factor Analysis**

The double-spondee words were broken down into single spondees for the analysis. Spondee words were composed of six phonemes in a pattern of consonant-vowel-consonant-consonant-vowel-consonant (CVCCVC) were included in the analysis. If a cluster of two or more consonants fell in one of the four /C/ places, they were scored for the correct identification of the unit instead of individual phonemes. The same conditions as indicated in Result showing the recognition of consonants and vowels were analyzed. The equation used in Boothroyd and Nittrouer (1988) was adopted to derive the j factor.

\[
j = \log(Pw)/\log(Pp),
\]

where \(Pw\) is the probability of recognizing the whole spondee word, and \(Pp\) is the probability of recognizing a part (i.e., one out of the six units composed the spondees). The \(j\) factors for the 2 NH listeners ranged from 1.21 to 2.33 in various conditions. The \(j\) factors for the 3 HI listeners ranged from 0.42 to 2.50. This preliminary observation may suggest a lower \(j\) factor for the HI listeners; however this needs to be tested with future data analysis.

**DISCUSSION**

The current study confirmed the co-occurrence of masking release and modulation interference in HI listeners when the redundant speech cues were degraded such that temporal envelopes were the more dominant cues. Masking release was observed more clearly when the modulation rate of the noise was 16 Hz, which is higher than the syllabic modulation rate, and when SNR’s were unfavorable. Modulation interference was seen when the noise was modulated at rates close to the syllabic modulation rate (4 Hz) and the SNR’s were favorable. This confirms and extends the previous findings of Nie et al., (2012), and Jin, Nie and Nelson (2013) for HI listeners.

In contrast, when the TFS and spectral details were impoverished for the NH listeners, the co-occurrence of masking release and modulation masking revealed a somewhat different pattern from that of HI listeners. Better speech understanding was seen when noise was amplitude modulated at both 16 and 4 Hz than when the noise was unmodulated, suggested masking release in both conditions. Overall, like for the HI listeners, speech understanding in the SAM-4Hz condition was somewhat less than that in the SAM-16Hz condition.

In quiet, both NH and HI listeners showed a benefit for 32 versus 8 channels of spectral information. However, in noise (steady and modulated), HI listeners showed no benefit of more channels. This suggests that broadened auditory filters have a marked effect on speech recognition performance in noise, but less in quiet.

Preliminary data analysis suggests that vocoding appears to affect consonant and vowel perception equally for NH listeners. However, for HI listeners, errors were more common for vocoded consonants than vowels.

The values of \(j\) factors in both HI and NH may reflect the strength of contextual cues that allow listeners to understand the spondaic words used in the current project. The possible smaller \(j\) factors measured in HI listeners may suggest that these listeners are using more contextual cues than NH listeners when temporal envelopes are the dominant cues as a result of the removal of TFS and the degradation of spectral resolution. Further analysis is ongoing.
ACKNOWLEDGMENTS

The authors appreciate Tess Koerner and Mary Richter for their time and dedication to the data analyses. Work supported by NIDCD 008306 to P.B.N.

REFERENCES


